Title: Privacy-Protecting, Reliable Response Data Discovery Using COVID-19 Patient Observations

Authors: Jihoon Kim[1†], Larissa Neumann[1,2,3†], Paulina Paul[1], Michael Aratow[4], Douglas S. Bell[5], Jason N. Doctor[6], Ludwig C. Hinske[2,3], Xiaoqian Jiang[7], Katherine K. Kim[8,9], Michael E. Matheny[10,11], Daniella Meeker[6,12], Mark J. Pletcher[13], Lisa M. Schilling[14], Spencer SooHoo[15], Hua Xu[7], Kai Zheng[16], Lucila Ohno-Machado[1,17]* for the R2D2 Consortium

Affiliations:

[1]UC San Diego Health Department of Biomedical Informatics, University of California San Diego, La Jolla, CA 92093, USA

[2]Institute for Medical Information Processing, Biometry, and Epidemiology (IBE), Ludwig Maximilian University of Munich, Munich, Bavaria 81377, Germany

[3]LMU Klinikum, Department of Anaesthesiology, Ludwig Maximilian University of Munich, Munich, Bavaria 81377, Germany

[4]San Mateo Medical Center, San Mateo, CA 94403, USA

[5]Biomedical Informatics Program, UCLA Clinical and Translational Science Institute (CTSI) Los Angeles, CA 90024, USA

[6]USC Schaeffer Center for Health Policy and Economics, Price School of Policy, University of Southern California, Los Angeles, CA 90031, USA

[7]School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA

[8]Betty Irene Moore School of Nursing, University of California, Davis, Sacramento, CA 95817, USA

[9]School of Medicine, Department of Public Health Sciences, Health Informatics Division, University of California Davis, Sacramento, CA 95817, USA

[10]GRECC Tennessee Valley Healthcare System, Nashville, TN 37212, USA.

[11]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN 37232, USA.

[12]Department of Preventive Medicine, Keck School of Medicine, Los Angeles, CA, USA

[13]Department of Epidemiology and Biostatistics, University of California, San Francisco, San Francisco, CA 94143, USA

[14] Data Science and Patient Value Program, University of Colorado Anschutz Medical Campus, Aurora, CO 80045, USA

[15]Division of Informatics, Department of Biomedical Sciences, Cedars Sinai Medical Center, Los Angeles, CA 90048, USA

[16]Department of Informatics, Donald Bren School of Information and Computer Sciences, University of California, Irvine, Irvine, CA 92697, USA

[17]Veteran Affairs San Diego Healthcare System, San Diego, CA 92161, USA

R2D2 Consortium authors are listed in the Supplementary Materials.

*Correspondence to: lohnomachado@health.ucsd.edu

†These authors contributed equally to this work.

One Sentence Summary: Publicly Sharing Knowledge on COVID19 Without Sharing Patient-Level Data: A Privacy-Protecting Multivariate Analysis Approach

Abstract: There is an urgent need to answer questions related to COVID-19's clinical course and associations with underlying conditions and health outcomes. Multi-center data are necessary to generate reliable answers, but centralizing data in a single repository is not always possible. Using a privacy-protecting strategy, we launched a public *Questions & Answers* web portal (https://covid19questions.org) with analyses of comorbidities, medications and laboratory tests using data from 202 hospitals (59,074 COVID-19 patients) in the USA and Germany. We find, for example, that 8.6% of hospitalizations in which the patient was not admitted to the ICU resulted in the patient returning to the hospital within seven days from discharge and that, when adjusted for age, mortality for hospitalized patients was not significantly different by gender or ethnicity.

Main Text: The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) pandemic represents a watershed event in public health and has highlighted numerous opportunities and needs in clinical and public health informatics infrastructure (*1*–*3*). One of the key challenges has been the rapid response of analyses and interpretation of observational data to inform clinical decision making and patient expectations, understanding, and perceptions (*4*–*8*).

Several initiatives are building COVID-19 registries or consortia to analyze electronic health record (EHR) data (*6*, *7*, *9*). The expectation is that these resources will provide researchers and clinicians access to a rich source of observational data to understand the clinical progression of COVID-19, to estimate the impact of therapies, and to make predictions regarding outcomes. Registries may contain limited data for patients diagnosed with COVID-19: the barriers for having more data are based on both privacy concerns and also on what elements have been deemed valuable by health professionals and researchers at a particular point in time. The problems with a new and evolving disease like COVID-19 is that we do not know what data or information will be most valuable. For example, in the pandemic's early stages, the dermatological and hematological findings were not evident, and those data were not included in registries or reports. Interest in specific laboratory markers (e.g., D-dimer, troponin) for these disturbances and additional symptoms (e.g. anosmia, conjunctivitis) has increased over time.

Additionally, it is challenging for researchers and clinicians to understand the structure and quality of the data in registries and other types of data repositories, and to formulate queries to consult the data in their institution. The process becomes more complicated when data from multiple institutions are involved.

Thus, the utilization of EHRs to characterize COVID-19 disease progression and outcomes is challenging. However, observational studies using EHR data may be useful when a research question does not lend itself to a randomized clinical trial (RCT). Observational studies may also help determine if results from RCTs replicate after relaxing eligibility criteria for real-world applications. While the scientific community has raised concerns about the reproducibility of findings, data provenance, and proper utilization of observational data, resulting in some COVID-19 articles being retracted (*10*), there remains a clear need to responsibly, ethically, and transparently analyze observational data to provide hypothesis generation and guidance in the pursuit of evidence-based healthcare.

We focus on using novel decentralized data governance and methods to analyze EHR-derived data. Researchers' questions posed in natural language are adapted to standardized queries using distributed data maintained in 12 health systems, covering 202 hospitals located in all U.S. states and two territories, and one international academic medical center (Table 1). This collaboration provides the capability to answer questions that require comparisons with historical data from over 45 million patients and uses a dynamic approach to account for an evolving awareness of the most impactful COVID-19 questions to answer and hypotheses to explore. Having access to complete EHR data from 10 of these health systems (two sites use COVID-19 registries), and not just a predefined list of key data elements, differentiates our approach from centralized registries and public health reports. The ability to build and evaluate multivariate models across a large number of health systems and to integrate results from registries differentiates our approach from most federated clinical data research network approaches.

Table 1. Participating sites. Cedars Sinai Medical Center (CSMC), University of Colorado Anschutz Medical Campus (CU-AMC), Ludwig Maximillian University of Munich (LMU), San Mateo Medical Center (SMMC), University of California (UC) Davis (UCD), Irvine (UCI), San Diego (UCSD), San Francisco (UCSF), University of Southern California (USC), University of Texas Health Science Center at Houston and Memorial Hermann Health System (UTH), Veterans Affairs Medical Center (VAMC). *Available data on hospital characteristics from 2018.

Institution	Hospitals	Beds	Discharges per year	EHR system	Data Source
CSMC	2	1,019	61,386	Epic	EHR
CU-AMC	12	1,829	106,325	Epic	EHR
LMU*	12	1,964	78,673	SAP/i.s.h.med QCare IMESO	COVID-19 Registry
SMMC	1	62	1,951	Harris Software (Pulsecheck) Cerner (Soarian) eClinicalworks	EHR
UCD	1	620	32,248	Epic	EHR
UCI	1	417	21,656	Epic	EHR
UCLA	2	786	47,491	Epic	EHR
UCSD	3	808	29,895	Epic	EHR
UCSF	3	796	48,120	Epic	EHR
USC	2	1,511	23,454	Cerner	EHR
UTH	17	4,164	233,890	Cerner	COVID-19 Registry
VAMC	146	13,000	676,402	ViSTa/CPRS	EHR
Total	202	26,976	1,361,491		

The responsibility of translating the question into code and of performing quality control processes lies among members of the Reliable Response Data Discovery for COVID-19 Clinical Consults using Patient Observations (R2D2) Consortium (see supplemental materials). The analyses do not require data transfer outside these institutions and reduce the risk of individual or institutional privacy breaches. We produce results that are publicly available as soon as they pass quality controls. In this approach, there is targeted analysis of specific data elements at a local level. Only the results of calculations (e.g., counts, statistics, coefficients, variance-covariance matrices) performed on data transformed into the Observational Medical Outcomes Partnership Common Data Model (OMOP CDM) from relevant patient cohorts are released from the healthcare institutions; no individual patient-level data are shared (*11*).

Between December 11, 2020 and August 31, 2020, our consortium had 928,255 tested patients for SARS-CoV-2, 59,074 diagnosed with COVID-19, with 19,022 hospitalized and 2,591 deceased. Our public Questions and Answers (Q&A) portal (https://covid19questions.org) provides answers to research questions using several univariate or multivariate analyses, including potential associations between mortality and comorbidities; pre-hospitalization use of anti-hypertensive medications; laboratory values and hospital events. For each question, we report on the number of participating institutions and the time period within which local queries were run. Figure 1 shows the graphical display of two answers.

Example 1. "*Many adult COVID-19 patients who were hospitalized did not get admitted to the ICU and were discharged alive. How many returned to the hospital within a week, either to the Emergency Room (ER) or for another hospital stay?*" The answer indicates 8.6%. This question is both important from the standpoint of understanding the natural course of disease and planning for needed resources. Although efforts are underway to understand post-discharge outcomes in COVID-19 infected patients, to date they have been limited to case series (*12*), modest sample sizes (*13*), or single-center or geographically concentrated health systems (*14*, *15*). These extant studies may also be hampered by fixed inclusion/exclusion criteria (*16*).

Example 2. "*Among adults hospitalized with COVID-19, how does the in-hospital mortality rate compare per subgroup (age, ethnicity, gender and race)?*" The answers from univariate and multivariate analyses (logistic regression, Fig. 2) indicate that *age* is a major risk factor, but *ethnicity* and *gender* are also significant when considered univariately. There is great interest and growing peer-reviewed literature on risk factors for COVID-19 mortality: the agility of our approach allows us to quickly re-run queries and rebuild models as new predictors become relevant and the understanding of the disease evolves (*17*–*20*).

Several other questions and answers are shown in the portal and are updated when answers are approved for posting. We also have a set of approved questions awaiting response from the majority of sites.

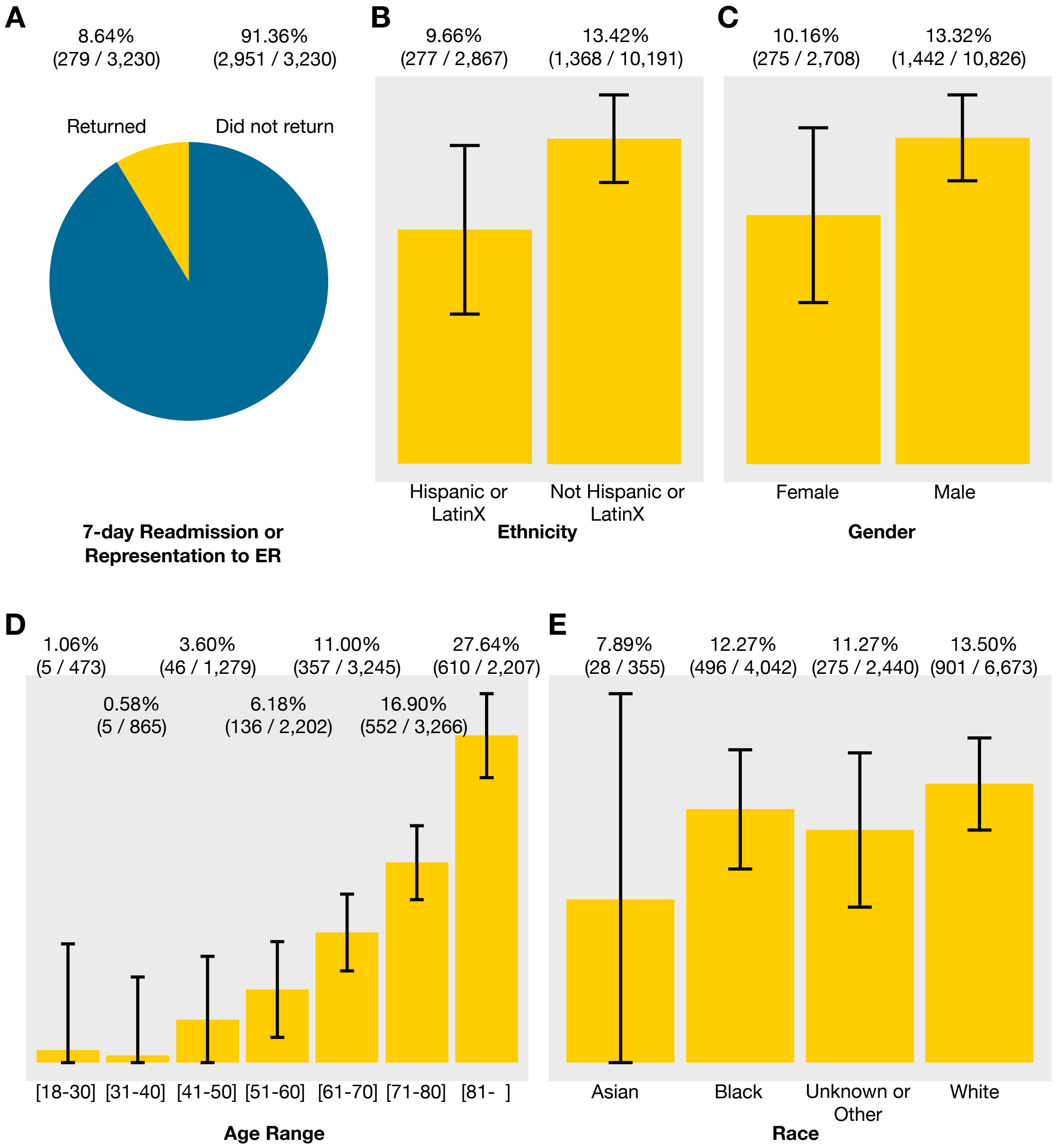

Fig 1. Examples of two COVID-19 *Questions and Answers*: Return to hospital and mortality. (A) 8.6% of hospitalizations without an ICU admission resulted in the patient presenting to the Emergency Room or a hospital readmission within seven days (data from ten health systems). (B-E) Unadjusted mortality rates from aggregated results are shown with 95% confidence intervals (data from ten health systems). Univariate analyses indicate that lower *age, Hispanic ethnicity,* and *female gender* are associated with lower mortality for adult hospitalized COVID-19 patients.

A

Variable	Coefficient	Standard Error	Z–statistic	P–value	Lower 95% CI	Upper 95% CI
Intercept	–6.572	0.508	–12.942	0.000	–7.567	–5.576
AGE	4.898	0.632	7.744	0.000	3.659	6.138
GENDER	0.290	0.182	1.591	0.112	–0.067	0.647
RACE_Asian	0.118	0.373	0.316	0.752	–0.613	0.849
RACE_Black	–0.739	0.437	–1.689	0.091	–1.596	0.119
RACE_UnknownOther	0.633	0.196	3.228	0.001	0.249	1.017
ETHNICITY	0.130	0.198	0.654	0.513	–0.259	0.518

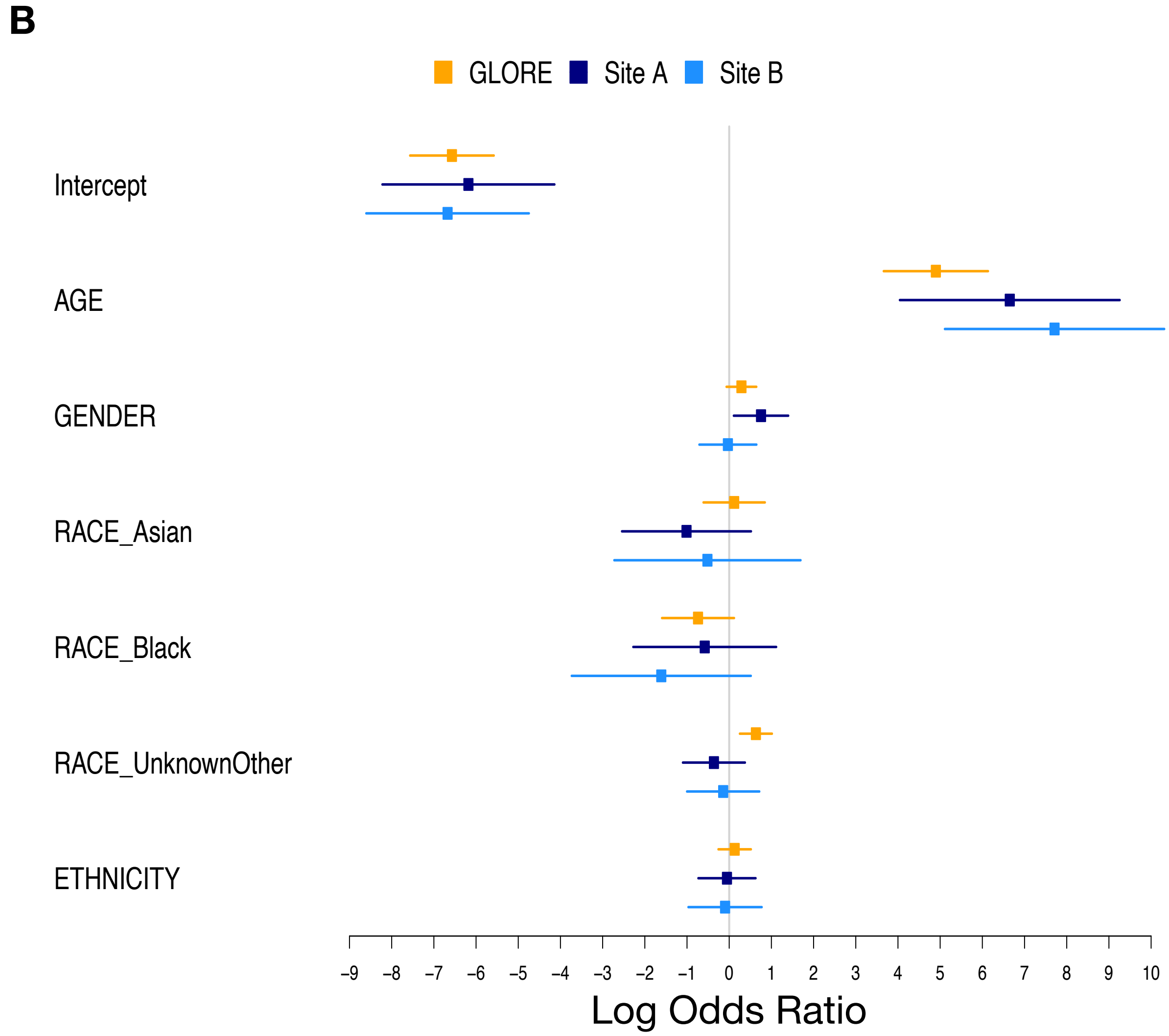

Fig 2. Regression Results. (A) Adjusted effects from the Grid Binary LOgistic REgression (GLORE) *(11)* federated logistic regression model (3,146 patients from eight health systems). The baselines were GENDER=*female*, RACE=*white*, ETHNICITY=*non-hispanic*. *Age* (in years) was divided by 100. After adjustment via distributed logistic regression, *age* remains significant. (B) Results from local logistic regression performed at two sites are also shown for comparison with GLORE results.

Institutions participating in our network are diverse in terms of organization, population served, prior investments in information technology, and location. A novel governance structure in our consortium allows us to distribute the workload across various teams without relying on a traditional coordinating center, instead including a Consortium Hub for certain functions.

This approach keeps patient data in-house, simplifies data use agreements, avoids delegation of control of patient data to another institution, and allows any institution to benchmark its results to those produced by the consortium, since all questions and respective final, aggregated answers, database query code, concept definitions and analytics code are made public. It complies with HIPAA, the Common Rule, the GDPR, and the California Consumer Privacy Act with regards to handling of patient data. Code sharing and public answers promote transparency and reproducibility without disclosing patient information or institutional information.

Our approach has advantages but also some limitations. The advantages are that we are able to, in relatively short time, publicly post answers to questions that are of general interest, using data from a spectrum of highly diverse institutions with different levels of information technology baselines and expertise in standardized data models and vocabularies, institutional policies, state and federal regulations. Because we keep data locally and only consult data elements that are necessary to answer specific questions, this approach has a lower risk of privacy breach when compared to registries in which patient data are exchanged or to distributed consortia in which summary-level results for each institution are reported. Additionally, since registries typically focus on a single disease or condition, they often lack comparator data from other patients, limiting the opportunity to characterize a new disease and discover how it differs from what we currently know. Participating sites do not need to transform all data into OMOP CDM and can decline to answer any questions they do not feel comfortable with or answer partially to ensure patient-level privacy by masking counts between 1 and 10. Institutional privacy is also preserved because all public answers combine the aggregate data from at least three Responding Sites (we do not specify which ones), but we keep all answers for audits. Making concept definitions, query code, and results available allows reproducibility and enables automated updates to the answers. A major advantage is that existing registries of consortia can serve as additional sites to help answer certain questions.

The limitations are inherent from considering all sites equal when formulating a final answer. Regional or institutional practice variations are not represented in the answers. Additionally, the distributed nature of the R2D2 consortium adds a requirement for a dynamic management team, the need to educate local leadership on distributed analytics, and potential for delays in certain decisions in order to exercise shared governance. A specific limitation of our current consortium is the preponderance of institutions based in California: eight out of 12 (67%), accounting for 17.5% of COVID-19 patients (Fig. 3). This was a convenience sample of organizations that had shared interests and a history of collaboration. We invite other institutions, consortia and registries worldwide to join us in answering questions of general interest.

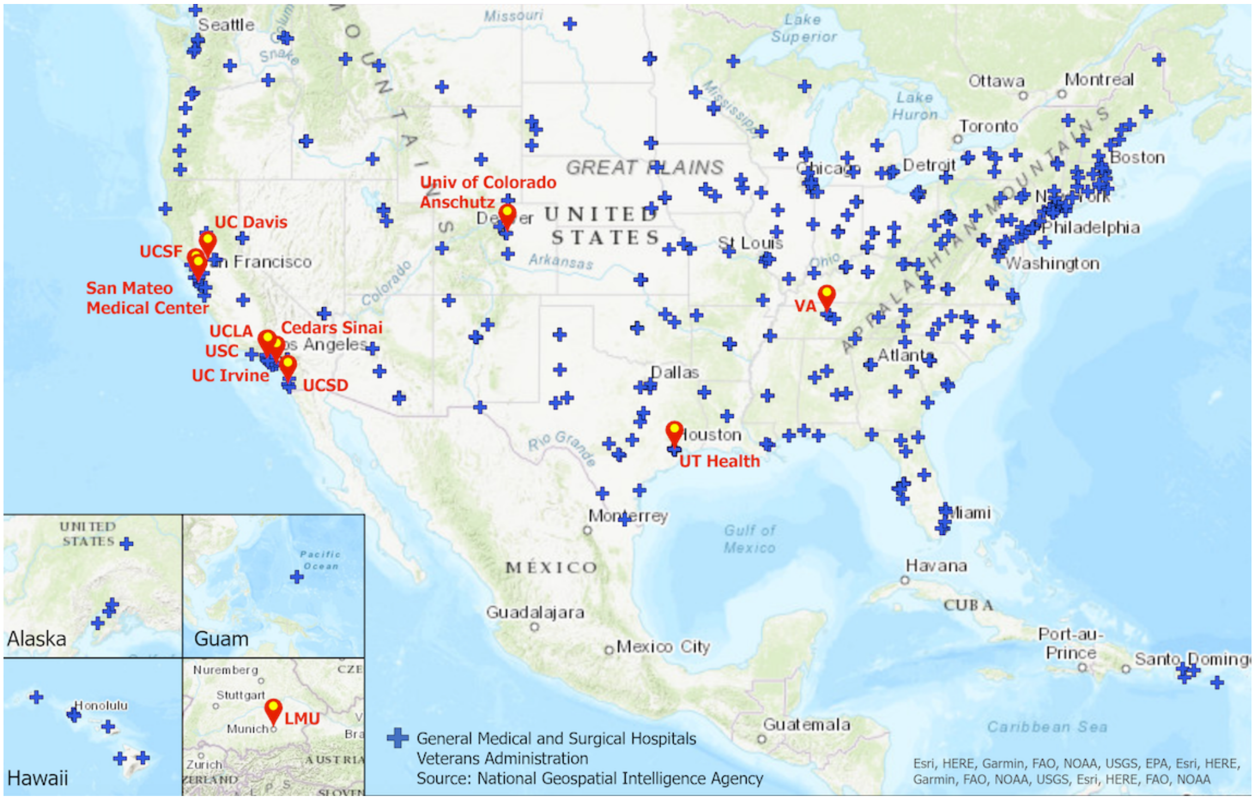

Fig 3. Location of consortium's medical centers and hospitals. Map by Ilya Zaslavsky

References and Notes:

1. N. Moradian, H. D. Ochs, C. Sedikies, M. R. Hamblin, C. A. Camargo, J. A. Martinez, J. D. Biamonte, M. Abdollahi, P. J. Torres, J. J. Nieto, S. Ogino, J. F. Seymour, A. Abraham, V. Cauda, S. Gupta, S. Ramakrishna, F. W. Sellke, A. Sorooshian, A. Wallace Hayes, M. Martinez-Urbistondo, M. Gupta, L. Azadbakht, A. Esmaillzadeh, R. Kelishadi, A. Esteghamati, Z. Emam-Djomeh, R. Majdzadeh, P. Palit, H. Badali, I. Rao, A. A. Saboury, L. Jagan Mohan Rao, H. Ahmadieh, A. Montazeri, G. P. Fadini, D. Pauly, S. Thomas, A. A. Moosavi-Movahed, A. Aghamohammadi, M. Behmanesh, V. Rahimi-Movaghar, S. Ghavami, R. Mehran, L. Q. Uddin, M. Von Herrath, B. Mobasher, N. Rezaei, The urgent need for integrated science to fight COVID-19 pandemic and beyond. *J. Transl. Med.* **18** (2020), p. 205.

2. D. F. Sittig, H. Singh, COVID-19 and the Need for a National Health Information Technology Infrastructure. *JAMA - J. Am. Med. Assoc.* (2020), , doi:10.1001/jama.2020.7239.

3. V. N. O'Reilly-Shah, K. R. Gentry, W. Van Cleve, S. M. Kendale, C. S. Jabaley, D. R. Long, The COVID-19 Pandemic Highlights Shortcomings in US Health Care Informatics Infrastructure: A Call to Action. *Anesth. Analg.* **131**, 340–344 (2020).

4. F. Eibensteiner, V. Ritschl, G. Ariceta, A. Jankauskiene, G. Klaus, F. Paglialonga, A. Edefonti, B. Ranchin, C. P. Schmitt, R. Shroff, C. J. Stefanidis, J. Vande Walle, E. Verrina, K. Vondrak, A. Zurowska, T. Stamm, C. Aufricht, Rapid response in the

COVID-19 pandemic: a Delphi study from the European Pediatric Dialysis Working Group. *Pediatr. Nephrol.* **35**, 1669–1678 (2020).

5. J. J. Reeves, H. M. Hollandsworth, F. J. Torriani, R. Taplitz, S. Abeles, M. Tai-Seale, M. Millen, B. J. Clay, C. A. Longhurst, Rapid response to COVID-19: Health informatics support for outbreak management in an academic health system. *J. Am. Med. Informatics Assoc.* **27**, 853–859 (2020).

6. D. A. Drew, L. H. Nguyen, C. J. Steves, C. Menni, M. Freydin, T. Varsavsky, C. H. Sudre, M. Jorge Cardoso, S. Ourselin, J. Wolf, T. D. Spector, A. T. Chan, Rapid implementation of mobile technology for real-time epidemiology of COVID-19. *Science (80-.).* **368**, 1362–1367 (2020).

7. A. T. Chan, D. A. Drew, L. H. Nguyen, A. D. Joshi, W. Ma, C. G. Guo, C. H. Lo, R. S. Mehta, S. Kwon, D. R. Sikavi, M. V. Magicheva-Gupta, Z. S. Fatehi, J. J. Flynn, B. M. Leonardo, C. M. Albert, G. Andreotti, L. E. Beane-Freeman, B. A. Balasubramanian, J. S. Brownstein, F. Bruinsma, A. N. Cowan, A. Deka, M. E. Ernst, J. C. Figueiredo, P. W. Franks, C. D. Gardner, I. M. Ghobrial, C. A. Haiman, J. E. Hall, S. L. Deming-Halverson, B. Kirpach, J. V. Lacey, L. Le Marchand, C. R. Marinac, M. E. Martinez, R. L. Milne, A. M. Murray, D. Nash, J. R. Palmer, A. V. Patel, L. Rosenberg, D. P. Sandler, S. V. Sharma, S. H. Schurman, L. R. Wilkens, J. E. Chavarro, A. H. Eliassen, J. E. Hart, J. H. Kang, K. C. Koenen, L. D. Kubzansky, L. A. Mucci, S. Ourselin, J. W. Rich-Edwards, M. Song, M. J. Stampfer, C. J. Steves, W. C. Willett, J. Wolf, T. Spector, The COronavirus Pandemic Epidemiology (COPE) Consortium: A Call to Action. *Cancer Epidemiol. Biomarkers Prev.* **29**, 1283–1289 (2020).

8. C. A. Longhurst, R. A. Harrington, N. H. Shah, A “green button” for using aggregate patient data at the point of care. *Health Aff.* **33**, 1229–1235 (2014).

9. M. Haendel, C. Chute, K. Gersing, The National COVID Cohort Collaborative (N3C): Rationale, Design, Infrastructure, and Deployment. *J. Am. Med. Informatics Assoc.* (2020), doi:10.1093/jamia/ocaa196.

10. M. R. Mehra, S. S. Desai, S. R. Kuy, T. D. Henry, A. N. Patel, Retraction: Cardiovascular disease, drug therapy, and mortality in Covid-19. *N. Engl. J. Med.* **382** (2020), p. 2582.

11. Y. Wu, X. Jiang, J. Kim, L. Ohno-Machado, Grid binary logistic regression (GLORE): Building shared models without sharing data. *J. Am. Med. Informatics Assoc.* **19**, 758–764 (2012).

12. X. Wang, Y. Zhou, N. Jiang, Q. Zhou, W. L. Ma, Persistence of intestinal SARS-CoV-2 infection in patients with COVID-19 leads to re-admission after pneumonia resolved. *Int. J. Infect. Dis.* **95**, 433–435 (2020).

13. J. Zheng, R. Zhou, F. Chen, G. Tang, K. Wu, F. Li, H. Liu, J. Lu, J. Zhou, Z. Yang, Y. Yuan, C. Lei, X. Wu, Incidence, clinical course and risk factor for recurrent PCR positivity in discharged COVID-19 patients in Guangzhou, China: A prospective cohort study. *PLoS Negl. Trop. Dis.* **14**, e0008648 (2020).

14. J. A. Lewnard, V. X. Liu, M. L. Jackson, M. A. Schmidt, B. L. Jewell, J. P. Flores, C. Jentz, G. R. Northrup, A. Mahmud, A. L. Reingold, M. Petersen, N. P. Jewell, S. Young, J. Bellows, Incidence, clinical outcomes, and transmission dynamics of severe coronavirus

disease 2019 in California and Washington: Prospective cohort study. *BMJ*. **369** (2020), doi:10.1136/bmj.m1923.

15. S. Richardson, J. S. Hirsch, M. Narasimhan, J. M. Crawford, T. McGinn, K. W. Davidson, and the Northwell COVID-19 Research Consortium, D. P. Barnaby, L. B. Becker, J. D. Chelico, S. L. Cohen, J. Cookingham, K. Coppa, M. A. Diefenbach, A. J. Dominello, J. Duer-Hefele, L. Falzon, J. Gitlin, N. Hajizadeh, T. G. Harvin, D. A. Hirschwerk, E. J. Kim, Z. M. Kozel, L. M. Marrast, J. N. Mogavero, G. A. Osorio, M. Qiu, T. P. Zanos, Presenting Characteristics, Comorbidities, and Outcomes Among 5700 Patients Hospitalized With COVID-19 in the New York City Area. *Jama*. **10022**, 1–8 (2020).

16. S. S. Somani, F. Richter, V. Fuster, J. K. De Freitas, N. Naik, K. Sigel, Mount Sinai COVID Informatics Center, E. P. Bottinger, M. A. Levin, Z. Fayad, A. C. Just, A. W. Charney, S. Zhao, B. S. Glicksberg, A. Lala, G. N. Nadkarni, Characterization of Patients Who Return to Hospital Following Discharge from Hospitalization for COVID-19. *J. Gen. Intern. Med.* (2020), doi:10.1007/s11606-020-06120-6.

17. J. A. Lewnard, V. X. Liu, M. L. Jackson, M. A. Schmidt, B. L. Jewell, J. P. Flores, C. Jentz, G. R. Northrup, A. Mahmud, A. L. Reingold, M. Petersen, N. P. Jewell, S. Young, J. Bellows, Incidence, clinical outcomes, and transmission dynamics of severe coronavirus disease 2019 in California and Washington: Prospective cohort study. *BMJ*. **369** (2020), doi:10.1136/bmj.m1923.

18. Y. Liu, X. Du, J. Chen, Y. Jin, L. Peng, H. H. X. Wang, M. Luo, L. Chen, Y. Zhao, Neutrophil-to-lymphocyte ratio as an independent risk factor for mortality in hospitalized patients with COVID-19. *J. Infect.* **81**, e6–e12 (2020).

19. N. Holman, P. Knighton, P. Kar, J. O'Keefe, M. Curley, A. Weaver, E. Barron, C. Bakhai, K. Khunti, N. J. Wareham, N. Sattar, B. Young, J. Valabhji, Risk factors for COVID-19-related mortality in people with type 1 and type 2 diabetes in England: a population-based cohort study. *Lancet Diabetes Endocrinol.* **0** (2020), doi:10.1016/s2213-8587(20)30271-0.

20. B. R. Yehia, A. Winegar, R. Fogel, M. Fakih, A. Ottenbacher, C. Jesser, A. Bufalino, R. H. Huang, J. Cacchione, Association of Race With Mortality Among Patients Hospitalized With Coronavirus Disease 2019 (COVID-19) at 92 US Hospitals. *JAMA Netw. open*. **3**, e2018039 (2020).

21. X. Dong, J. Li, E. Soysal, J. Bian, S. L. DuVall, E. Hanchrow, H. Liu, K. E. Lynch, M. Matheny, K. Natarajan, L. Ohno-Machado, S. Pakhomov, R. M. Reeves, A. M. Sitapati, S. Abhyankar, T. Cullen, J. Deckard, X. Jiang, R. Murphy, H. Xu, COVID-19 TestNorm: A tool to normalize COVID-19 testing names to LOINC codes. *J. Am. Med. Informatics Assoc.* (2020), doi:10.1093/jamia/ocaa145.

22. N. G. Weiskopf, S. Bakken, G. Hripcsak, C. Weng, A Data Quality Assessment Guideline for Electronic Health Record Data Reuse. *eGEMs (Generating Evid. Methods to Improv. patient outcomes)*. **5**, 14 (2017).

23. Centers for Disease Control and Prevention, ICD-10-CM Official Coding and Reporting Guidelines, April 1, 2020 through September 30, 2020. **19**, 19–21 (2020).

24. J. C. E. Lane, J. Weaver, K. Kostka, T. Duarte-Salles, M. T. F. Abrahao, H. Alghoul, O. Alser, T. M. Alshammari, P. Biedermann, J. M. Banda, E. Burn, P. Casajust, M. M.

Conover, A. C. Culhane, A. Davydov, S. L. DuVall, D. Dymshyts, S. Fernandez-Bertolin, K. Fišter, J. Hardin, L. Hester, G. Hripcsak, B. S. Kaas-Hansen, S. Kent, S. Khosla, S. Kolovos, C. G. Lambert, J. van der Lei, K. E. Lynch, R. Makadia, A. V Margulis, M. E. Matheny, P. Mehta, D. R. Morales, H. Morgan-Stewart, M. Mosseveld, D. Newby, F. Nyberg, A. Ostropolets, R. W. Park, A. Prats-Uribe, G. A. Rao, C. Reich, J. Reps, P. Rijnbeek, S. M. K. Sathappan, M. Schuemie, S. Seager, A. G. Sena, A. Shoaibi, M. Spotnitz, M. A. Suchard, C. O. Torre, D. Vizcaya, H. Wen, M. de Wilde, J. Xie, S. C. You, L. Zhang, O. Zhuk, P. Ryan, D. Prieto-Alhambra, Risk of hydroxychloroquine alone and in combination with azithromycin in the treatment of rheumatoid arthritis: a multinational, retrospective study. *Lancet Rheumatol.* (2020), doi:10.1016/s2665-9913(20)30276-9.

Acknowledgments: We thank Dr. Michael Hogarth for insightful comments.

Funding: Funded by the Betty and Gordon Moore Foundation #9639. The distributed analytics algorithm was funded by NIH-R01GM118609. Trainees were funded by NIH-T15LM011271. LN was funded by DIFUTURE (BMBF grant 01ZZ1804C); **Author contributions:** JK, LN and LO-M drafted the manuscript. PP, MA, DB, JD, LH, KK, DM, MM, MP, LS, SS, HX, KZ organized local teams for answering questions. All authors reviewed the manuscript. LO-M supervised the study and acquired the funding. See supplemental materials for a complete list of consortium authors; **Competing interests:** Hua Xu has financial interest at Melax Technologies Inc. Other authors declare no competing interests; and **Data and materials availability:** Answers are available at https://covid19questions.org and code is in https://github.com/DBMI/R2D2-Public.

Supplementary Materials:

Materials and Methods
Figs. S1 to S5
Tables S1 to S5
R2D2 Consortium Members